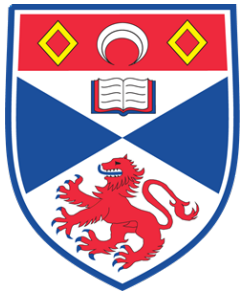
University of St Andrews

CS5014 Coursework 1

Machine Learning

Author:
150008022

February 16, 2019



Goal

The goal of this practical was to cleanse and process real world data in order to produce a regression model, and evaluate its performance. [1]

Part I
Loading and Cleaning the Data

Numpy was used to load the csv file. The header row was skipped, and to ensure there were no missing values, the invalid raise flag was also used when parsing the data. This would raise an exception if any rows were found to be missing data.

The input and output columns were separated into two variables x and y, as in accordance with the notation used in lectures.

Part II
Analysing and Visualising the Data

Firstly, a histogram of each of the input variables was plotted in order to visualise the distribution of the values. When comparing to the histograms from the given paper TODO ref paper, most of the plots matched. Since only 10 bins were used each time (the default if not specified by numpy.hist), and the paper would sometimes use more, the graphs did not exactly match. However, it was still clear that none of the variables had a gaussian distribution.

Conclusion

References

[1] Facebook. create-react-app.